Restricted Directed Information for Gene Regulatory Interactions Inference

Arman Rahimzamani*, Xiaojie Qiu†, Cole Trapnell†, Sreeram Kannan*

*Department of Electrical and Computer Engineering, University of Washington, † Department of Genome Sciences, University of Washington

Gene Regulatory Network Inference

Most biological processes, either in development or disease progression are governed by complex *gene regulatory networks*. Hence inferring these networks is of vital importance to biologists. However there are many challenges here to tackle, including but not limited to distinguishing upstream regulatory genes from their targets directly downstream.

Single-cell Transcriptome Seq Experiments

Single-cell transcriptome sequencing experiments (scRNAseq) have attracted the attention of algorithm developers working on gene regulatory network inference for two reasons. First, scRNAseq experiments now routine produce thousands of independent measurements may open the door to sufficiently powered inference. Second, algorithms that order cells along "trajectories" that describe development or disease progress offer a tremendously high "pseudotemporal" view of gene expression kinetics.

$\longrightarrow$ One might think of inferring these regeulatory relationships via the methods tailored for temporal data.

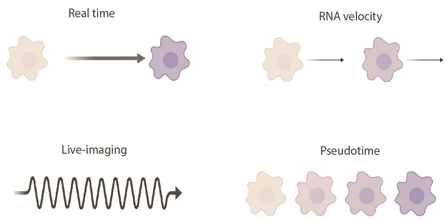

Figure 1: Various Transcriptomic Temporal Experiments.

Scribe Package

- A *Information Theoretic* based causal inference method from temporal data
- Agnostic to the particular measurement technology used in the experiment
- Incorporates a *visualization framework* to visualize 1) The response function, 2) Causal interaction as well as 3) combinatorial regulation between gene pairs

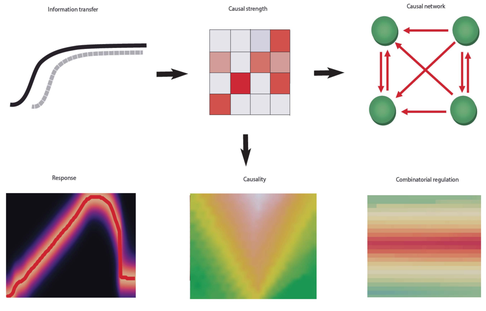

Figure 2: Scribe in a Nutshell.

What's inside Scribe?

- Restricted Directed Information (RDI):

$$RDI\left(X \to Y|Z\right) = I\left(X(t-1); Y(t)|Y(t-1), Z(t-1)\right)$$

- Uniform Restricted Directed Information (uRDI):
In RDI, we replace the ordinary Conditional Mutual Information (CMI) with uniform CMI:

$$uCMI_{X\to Y|Z}(\mathbb{P}_{Y|X,Z}) := CMI_{X\leftrightarrow Y|Z}(\mathbb{U}_{X,Z}\mathbb{P}_{Y|X,Z})$$

- Estimators implemented: KSG-based KNN estimators:

- Proven to be consistent.
- Their finite-sample performance is shown to be superior to that of the alternative estimators.

- The Scheme of the overall algorithm:

1. Calculate the *unconditioned* pairwise RDI/uRDI for every pair of genes (i, j).
2. For each pair (i, j), find the L incoming genes to j with the highest RDI values excluding i itself.
3. Recalculate the pairwise RDI/uRDI for (i, j) *conditioned* on the L genes above.

Scribe Applied to Synthetic Data

- Emulated a 13-gene nonlinear Neurogenesis system.
- All four Real-time, Live-imaging, RNA velocity and pseudo-time data created from the system.

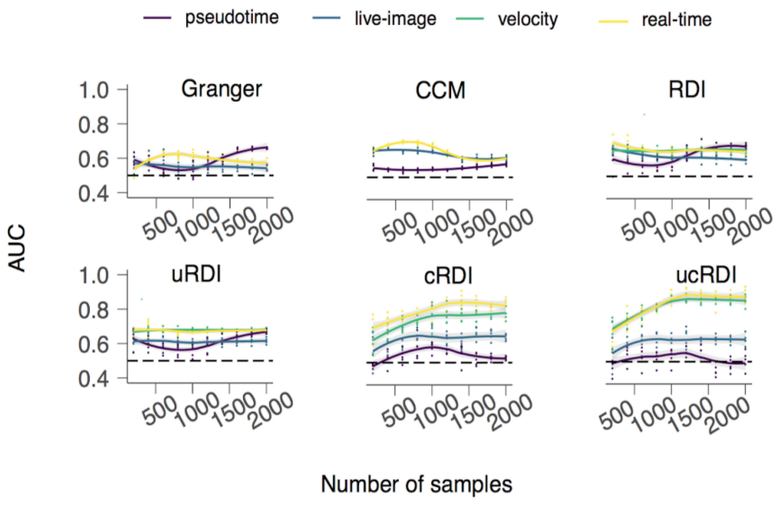

Figure 3: Results on Neurogenesis data.

Scribe Applied to Real Data

- Recovering Myoplesis Network:

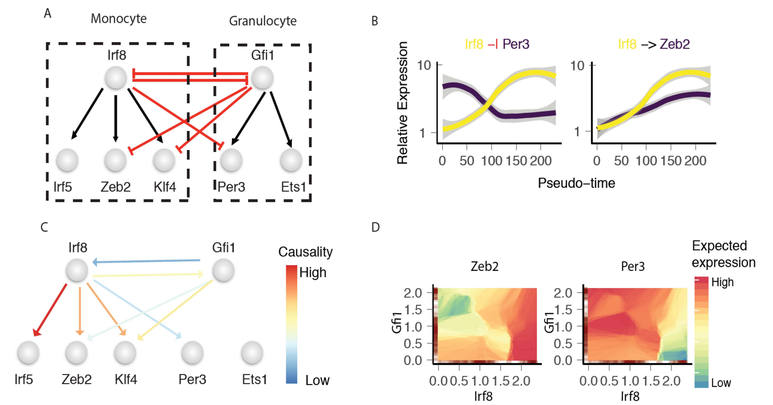

Figure 4: Scribe recovers a core regulatory network responsible for myelopoiesis. (A) A core network describes key regulators during the specification of monocytes and granulocytes based on data collected from perturbation experiments, bulk ATAC-seq and ChIP-seq data. (B) Examples of gene-target pair kinetic curves over pseudotime along the monocyte lineage. (C) Scribe infers the expected core regulatory network interactions for myelopoiesis. Only significant edges are plotted while color of edge corresponds to the causality score. (D) Visualization of combinatorial gene regulation from Irf8 and Gfi1 to Zeb2 or Per3.

- Inference from the Velocity Data:

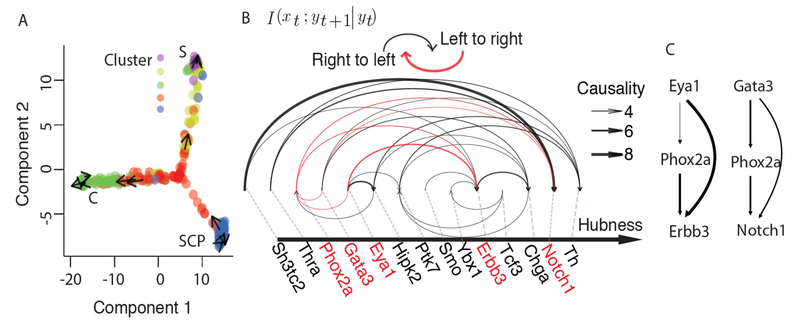

Figure 5: Network inference in Scribe with RNA-velocity. (A) RNA-velocity vector projected onto the first two latent dimensions. S: Sympathoblasts; C: Chromaffin. SCP: Schwann Cell Progenitor. (B) A core regulatory network for chromaffin cell commitment inferred based on RNA-velocity. The width of the edge corresponds to the normalized causality score returned after applying CLR regularization on RDI values. (C) Two potential coherent FFL (feed-forward loop) motifs of chromaffin differentiation are discovered from the core network. Edge width corresponds to the regulation strength.

Comments

- Gene regulatory network inference from observational measurements is widely regarded as amongst the most difficult problems.
- The quality of the data collection experiments needs to be improved: Noise and Dropouts are still big challenges.
- Lack of temporal/spatial coupling between gene measurements drastically hurst the quality of inference.